

# 记忆的结构

甄景贤 (King-Yin Yan) and Joseph Cheng

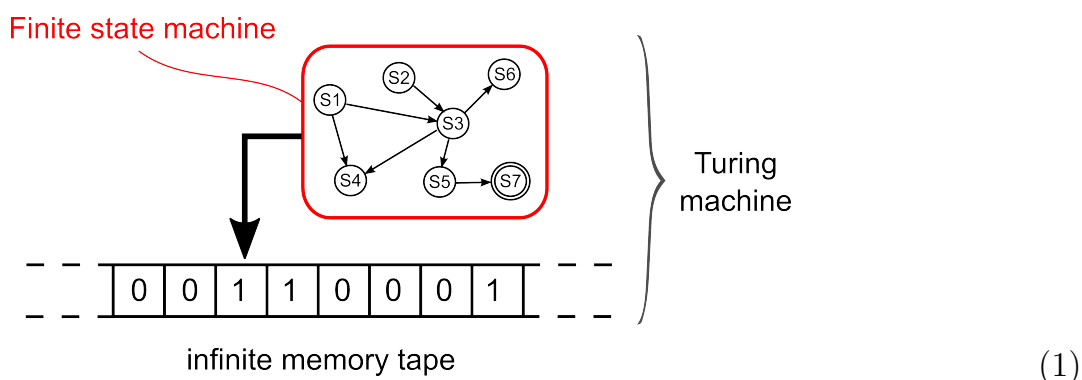
General.Intelligence@Gmail.com

**Abstract.** (This is a draft...)

智能系统中记忆的设计似乎是一个比较 open-ended question。暂时我列出一些关于记忆的点子，但仍未综合成一个系统的理论....

## 0 Turing machine 的启示

深度学习在近年很火，但 deep NN 的缺点是没有记忆。有限自动机 (finite state machines) 不是全能的计算器，它和 Turing machine 的分别就是缺少了那条「记忆磁带」：



所以，深度神经网络 + 记忆 就可以变成 universal 的计算器。如何设计「可微分」的记忆是一个重要课题，因为可微分的记忆可以用 gradient descent 学习。

## 1 什么是 episodic memory?

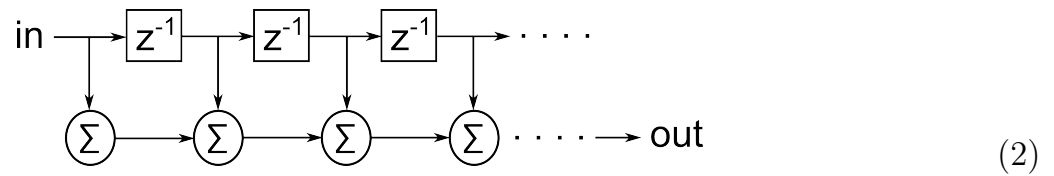
在 [?] 中我们提出了智能系统的 minimal architecture，但是它没有对历史的记忆。换句话说，只能留意当下发生的事件，但不能记住一段故事。

这牵涉到「什么是记忆？」的问题。在 minimal architecture 里， $F$  代表 “static knowledge”，亦即（相对地）永恒不变的知识 / 规律，而  $x$  代表当下的状态 / 短期记忆，亦即 “dynamic knowledge”。Episodic memory 介乎「长期」与「短期」记忆之间。

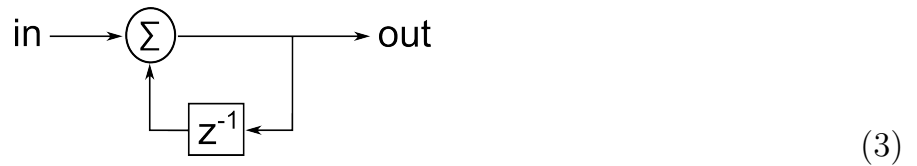
换句话说，最简单的 reactive system 其实是一个只有瞬时记忆和永恒记忆的系统。例如无人驾驶车，它的瞬时记忆就是路面状态，而永恒记忆是「转弯、打灯、泊车」等动作。它根据路面状况（例如有没有撞车等奖励 / 惩罚），学习永恒记忆中的 responses。

我们的目标是要设计有 **limited memory** 的系统，令系统能够根据 episodic memory 来学习。

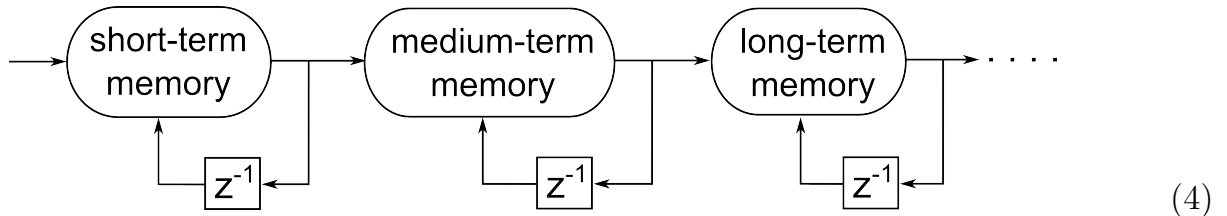
在 信息处理 (signal processing) 理论中， $z$ -transform 以  $z^{-1}$  可以代表 时间延迟 1 步 (1-step time delay unit)。用一连串的  $z^{-1}$  可以组成长度为  $k$  单位的记忆：



以上的无穷串列可以用这个 recursive 结构代替：



多层的 hierarchical 记忆或许可以这样实现：



$Z$ -transform 的连续时间版本是 Laplace transform （不知有没有用？）

上面的记忆结构来自 Simon Haykin 的经典著作《Signals and Systems》[1]。

## 2 Mental state $x$ = working memory = 命题的集合

在第二篇论文 [2] 的 §2.2.1 里讲述了思维状态  $x$  应该被分拆成若干个命题的集合。 $x$  是  $M$  个 thoughts 的集合， $M$  是 working memory 的大小。认知科学里有个说法 [?]:

$$\text{size of human working memory} \approx 7 \pm 2 \text{ items} \quad (5)$$

但这些 items 可以有 **chunking** [?], 例如去超级市场买东西, 「意大利粉、茄汁、芝士、香肠」这 4 件东西可以聚合成一个 chunk, 这样可以记住的 items 数目多很多。

## 3 记忆 = 状态的历史

现在回看状态方程：

$$\boxed{\text{连续时间}} \quad \dot{x} = f(x) \quad (6)$$

$\dot{x}$  是状态  $x$  的改变方向，换句话说，这是描述状态变化的方程。但状态的变化  $f(x)$  不取决于状态的历史 ( $x = x|_t$  仅代表状态在时间  $t$  的值)，所以这个系统没有记忆。如果想要记忆的话，一个简单的做法是令：

$$\dot{x} = f(x|_t, \overbrace{x(t \in \mathbb{T})}^{\text{状态 } x \text{ 的历史}}) \quad (7)$$

$f$  就是我们的神经网络。 $f$  的输入是：

- $x|_t$  = 当下的状态 = 一些命题的集合；命题没有 time stamp
- $x(t \in \mathbb{T})$  = 状态的历史；要考虑每个状态的 time stamp

## 4 联想记忆 = associative / content-addressable memory

如果用 content-addressable 的方法，似乎可以不用 time stamps，只要每个记忆 items 之间用「时间先后-link」连接就可以。人脑的记忆似乎是这样的。

## 5 Convolution and Fourier transform

爱因斯坦说时间是第 4 度空间，从物理学的角度看，时空问题可以 reduce 成类似 3 度空间的处理，亦即是机器视觉的问题，这问题基本上已经解决了。

对 3D 图像处理的方法，是用 convolution，亦即是用一个内核 (kernel) 不断重复辨认空间中的某一局部、到另一局部、and so on....。这种做法能做到资讯压缩，其原因是 weight-sharing，亦即是说，用一个内核的 weights，代表了原本要用很多神经元的 weights，这就达到压缩的效果。

「4 维时空」的做法应该是类似的。

根据 convolution theorem[?]:

$$\text{convolution in time domain} = \text{multiplication in frequency domain} \quad (8)$$

换句话说，如果在空间上做了 Fourier transform (或许包括 wavelet transform?) 其作用等于在时间上做 convolution。

在一个智能系统里，「时间」的地位和「空间」可能有不平等，因为智能系统是在时间上做行动的，它的记忆结构可能是根据时间而 organize，这样导致分析的不方便。在我现时的 design 中，时间有特殊地位。

或许将时空看成 4D 会带来理论上的简洁，也就是处理上更方便....?

如果 sensory data 就是一大块 4D 资料，智能的问题就是 unsupervised pattern recognition in 4D，在理论上这很简单。

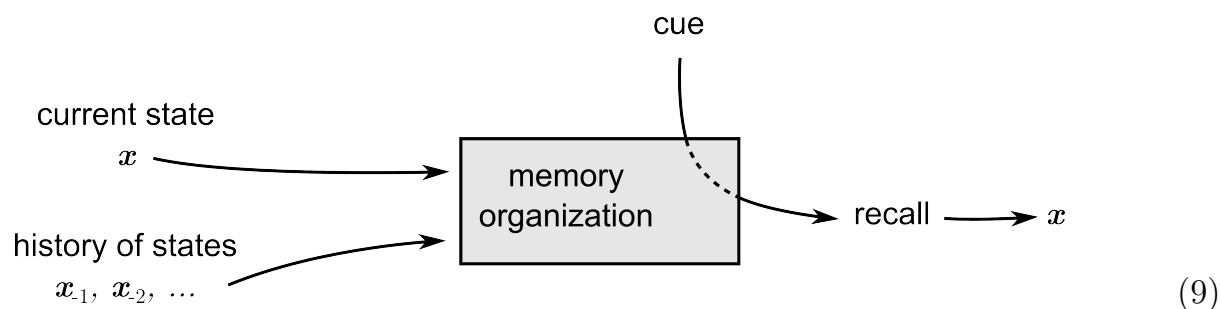
但是这种做法忽略了智能系统和环境之间的互动，包括奖励等。例如智能系统可以「玩弄」某些物件去了解它的特性，或者和人类交谈、问问题等。所以有必要使用 reinforcement learning (RL) 的框架。

在 RL 框架内，「当下的状态」 $x$  是时间的函数。RL 理论是相对於 state space trajectory 而发展的。但「4D 影片」是状态  $x$  在时间上累积而成的。

## 6 Desiderata

- able to recognize (辨识) patterns within time and across time
- cue-based retrieval (或者说，基於 **attention** 的记忆读取)
- mental state 分拆为一束命题的集合
- 遗忘的机制，with graceful degradation
- **hierarchical** 组织 (基於时间或其他 attributes)
- 

## 7 简单设计



- 记忆应该如何组织?
- retrieval 根据的是什么?  
应该是 relevance。注意: rules 在 **KB** 里, facts 在 memory 里。这个 config 要保持吗?
- 

## References

1. Simon Haykin and Barry van Veen. *Signals and systems*. John Wiley & Sons, 2003.
2. King Yin Yan. A bridge between logic and neural. (to be submitted AGI-2017).